UNIX Today!

087 DECEMBER 9, 1991 The newspaper Of Open Systems Computing A CMP Publication®

The Promise Of The WAIS Protocol

Emerging Standard Represents First Step Toward Unifying Data Search & Retrieval

BY JASON LEVITT

It doesn't take an expert to see that the state of moder information handling is neither open nor unified. A trip to the main library at the University of Texas at Austin–one of the top 10 college library systems in the U.S.–confirms this.

The primary card catalog is contained on an IBM mainframe accessible through various synchronous block-mode terminals scattered about the main library, and also accessible via modem through a rather crude dial-up facility.

In the main reference room, an OCLC (On-line Computer Library Center) terminal allows access to other university card catalogs; several IBM PCs are available to search CD-ROMs for bibliographical citations and abstracts on a variety of subjects; and a LEXIS/NEXUS terminal can be used for researching major U.S. court decisions. In the engineering library, an IBM PC with CD-ROM is available for searching U.S. patents.

If one were to compare information accessibility at this facility to computer resource accessibility, things here are still in the early 1980s or late '70s. Each of the systems mentioned are primarily standalone and proprietary, having their own information retrieval and organizational formats with little, if any, interoperability between databases.

While the monster mainframe card catalog might provide pointers to many sources, it is ignorant of most other on-line sources and almost never provides the most current information on subjects, despite the best efforts of its administrators. What these information-handling systems need is a dose of open systems standards and technology, the same technology that is changing the face of modern computing.

Enter WAIS, for Wide-Area Information Server, a fledgling step in the overwhelming effort needed to unify information search and retrieval technology. WAIS is an emerging open systems standard protocol for query and retrieval of information. WAIS, pronounced "ways," is the brainchild of Brewster Kahle, an employee of Thinking Machines Corp. (TMC), the No. 2

WAIS Server Source File

```
( :source
  :version 3
  :ip-name "nextbox.utoday.com"
  :tcp-port 5001
  :database-name "UT_TECH"
  :cost 0.00
  :cost-unit :free
  :maintainer "jason@nextbox.utoday.com"
  :description "Server created with WAIS release
8 b2 on Mon Nov 18 16:54:19 1991 by
jason@nextbox.utoday.com
UNIX Today! technology articles by Jason Levitt
The files of type text used in the index were:
  /LocalLibrary/WAIS/articles/ABCstory.txt
  /LocalLibrary/WAIS/articles/AIX3.1FS.txt
  /LocalLibrary/WAIS/articles/Benchinfo.txt
  /LocalLibrary/WAIS/articles/LPFstory.txt
  /LocalLibrary/WAIS/articles/MacXstory.txt
  /LocalLibrary/WAIS/articles/Solbourne.txt
  /LocalLibrary/WAIS/articles/SunStory.txt
  /LocalLibrary/WAIS/articles/XSerialArticle.txt
  /LocalLibrary/WAIS/articles/Xarticle.txt
  /LocalLibrary/WAIS/articles/Xcontrib.txt
"
)
```

Figure 1

supercompute manufacturer, behind Cray, and purveyor of fine, massively parallel systems.

The basis for WAIS is the rapidly growing electronic-publishing movement, which is seeing more and more materials, usually available only in book form, "published" or placed onto electronic media such as disk and tape, where it can be accessed with a computer.

WAIS TECHNOLOGY

WAIS is a protocol for the transmission of query and retrieval information, much like the information you would use to search a library card catalog. It is, in fact, an extension to an existing protocol standard called Z39.50, the Information Retrieval Service Definitions and Protocol Specification for Library Applications.

The Z39.50 standard was created by a group called NISO, the National Information Standards Organization, and is designed for use in electronic library card catalogs. Z39.50 essentially specifies formats for search requests directed at a database and formats for document retrieval requests. WAIS extends the Z39.50 standard to allow, among other things, discrete portions of documents, called "chunks," to be retrieved. This is especially useful in low-bandwidth situations such as serial links, where transferring an entire document in response to a query would be prohibitively time-consuming.

☆ Text ☆ Retrieval

The WAIS protocol fits neatly at the top of the ISO 7-layer protocol model at the application and presentation layers. This makes it extremely portable to differing network environments such as TCP/IP and X.25.

Like any good open standard, the WAIS protocol does not specify or limit the technology at either end of the wire. A WAIS client can be as simple as a command line interface that takes a database name, network address and query string as input, or as complex as a combination spreadsheet and database that constantly updates in real time, based on client/server activity taking place in the background. The only condition is that the client and server exchange query and retrieval information using the WAIS protocol.

The free WAIS source code, discussed later, implements a very typical client/server model for Unix-based Internet applications. The server creates and waits on a socket attached to a well-known port. Clients attach to the port using the port number and network address of the machine. The server accepts a request, forks a child process to handle the request, and then continues to wait and service other requests.

Requests for information are largely governed by special text files maintained by the WAIS server, called "sources," that vaguely resemble library catalog cards. Figure 1 shows a source I created containing 10 of my previous technology articles for *UNIX Today!* There is enough information in the source structure, network address, TCP port number and database name for any other machine on the network running a WAIS client to locate, understand and access the information in the database.

Not surprisingly, WAIS is already being used to connect archive sites on the Internet running on various Unix-based machines as well as proprietary systems such as Macintosh and NeXT. According to Brewster Kahle, there are approximately 80 sites running public WAIS servers and many more running WAIS privately within corporations and academia. A FidoNet WAIS server site was recently added to this collection of public sites running SLIP over a 9,600-bps serial link.

General WAIS Information

Thinking Machines Corp.
1010 El Camino Real, Ste. 310
Menlo Park, CA 94025
415-329-9300 Fax: 415-329-9329

Bibliography of available WAIS documents.
Send electronic mail to: barbara@think.com

Accessing a WAIS client on the Internet

Telnet to quake.think.com, login as **wais**

Getting involved with the Nat'l Public Network

Electronic Frontier Foundation
155 Second Street
Cambridge, MA 02141
617-864-0665
E-mail: eff@eff.org

FREE WAIS SOFTWARE

I like software that you can use to get some meaningful work done quickly without having to dig too deeply into documentation. The freely available WAIS software fits that description. In the

UNIX Today! labs, I decided to put together a small heterogeneous network and run WAIS.

Acting as the WAIS server system (and also a client) was a NeXTstation. Attached over Ethernet was a Macintosh running MacOS and a Sun 3/60 running SunOS 4.1. The free WAIS software included NeXT and Mac binaries and complete source code for the Unix systems, in this case the Sun. I dug out my archives of personal Unix electronic mail, about 10 Mbytes' worth, and used the indexing program included with the WAIS server to create a hashed database. I did the same with 10 of my old technology articles written for *UNIX Today!* The databases, or "sources," are listed in Figure 2.

The WAIS indexing program knows about the format of many common types of structured on-line data such as electronic mail, *netnews*, PICT/GIF/TIFF files and biology abstract formats, and it also handles straight ASCII text.

There was also a database of WAIS documentation, created automatically by the server program, and a directory of all sources I created called "directory of information" that simply points to all the databases. After creating the databases, I ran the WAIS server program, called *waisserver*, on the NeXTstation, which sits and waits for incoming WAIS client requests.

Once the *waisserver* was running, I could access it using the clients, called WAIS stations. On the Sun, which was running X/Motif, I chose to use the Motif client. I also used the Mac and NeXT WAISstations. In order to access a *waiserver*, I first had to set up my sources. Figure 3 shows a source setup window for the Mac client. I had named my database of articles "UT-TECH" on the WAIS server. The access method, "Contact," was MacTCP, Apple's TCP/IP protocol stack.

As shown in Figure 2, I decided to search my mail archives and technology articles for references to NCD's Xremote protocol. The results appear in the scrolling list. If the result is an entire file, such as the article contained in the file "XSerialArticle.txt," the path name for the file is listed after it.

The other results in the list are individual E-mail messages that actually are in several large text files on the WAIS server. Because the WAIS indexing program understands E-mail format, it was able to index individual E-mail messages in my E-mail archive files and transfer only those E-mail messages pertinent to the client query.

By clicking on a document in the Results window, the portion of the result most relevant to my query appears in another window. The *waisserver* uses a simplistic approach to interpreting my request for information about Xremote. It looks for the word "xremote"–the search is case-insensitive–in mail messages and headers and displays matching documents and mail messages in the results window. This turns out to be adequate as long as you put meaningful words in your query.

TMC has a much more sophisticated searching mechanism in its Internet server, *quake.think.com;* however, the search source code is not freely available.

One of the key features of the WAIS protocol is its ability to allow secondary search criteria. In Figure 2, the criteria would be entered by copying a result, or chunk of a result, to the "which are similar to" window. A subsequent search would use any words contained in that window as additional search criteria. Repeatedly using that method can quickly refine the search parameters.

On-Line WAIS Discussions And Development

alt.wais newsgroup on USENET

Join mailing lists by sending e-mail to:

wais-discussion-request@think.com – Weekly digest of mail from users and developers

wais-interest-request@think.com – Infrequent announcements of new releases

wais-talk-request@think.com – Developers' mailing list

Free WAIS client software

Clients for NeXT, X, Macintosh, Unix ASCII, GNU Emacs and Motif.
Anonymous FTP to think.com in the directory /wais

Clients for VMS, MS-DOS, Novell LAN Workplace and SunView.
Anonymous FTP to samba.oit.unc.edu in the directory /pub/wais/UNC

Free WAIS server software

Servers for NeXT and various Unix platforms
Anonymous FTP to think.com in the directory /wais

AN OPEN END

The next version of the WAIS protocol should be officially folded into the Z39.50 standard this month and is expected to include multimedia support and integral support for English-language queries. These enhancements should add considerable clout to WAIS, given the infant state

of commercial multimedia query/retrieval technology.

WAIS software is freely available from a number of sites. Unfortunately, the WAIS client program can only be obtained via anonymous FTP at this time, which means you have to have direct Internet access.

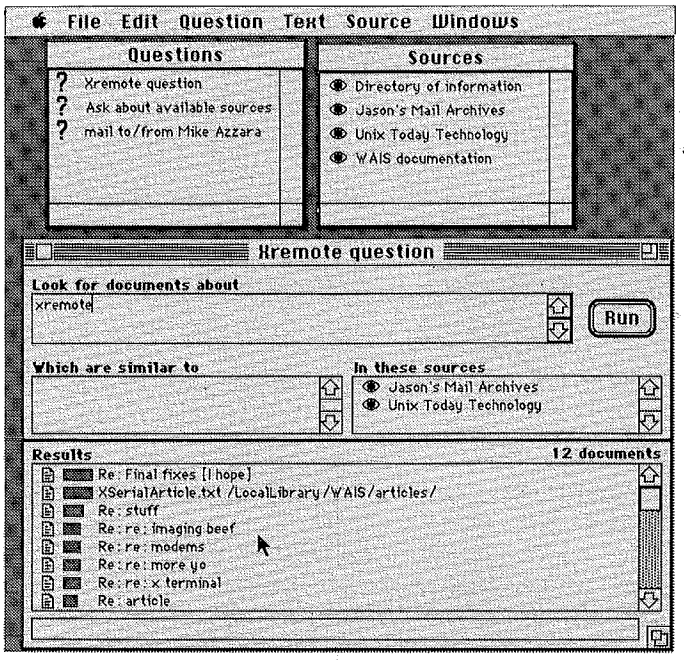

Figure 2: Mac WAIS client shown with results of a search for "xremote"

The WAIS server and X-based client program for Unix are available on *uunet.uu.net* in the directory */networking/distrib-is/wais.*

My small network experiment with WAIS only touched on its full potential; however, for my small database needs, it was quite useful. The free WAIS software is, like the MIT X software, meant as reference software for further development, not as a commercial-quality implementation.

I encountered bugs, such as a persistent permissions error from the NeXT client, and strange window clipping from the Motif client, and I have yet to get the *waisserver* running cleanly under SVR4. But when the software is open and free, who cares?

The vision of WAIS is not only easy access, retrieval and publishing of information, but the creation of a marketplace that can encourage new information sources.

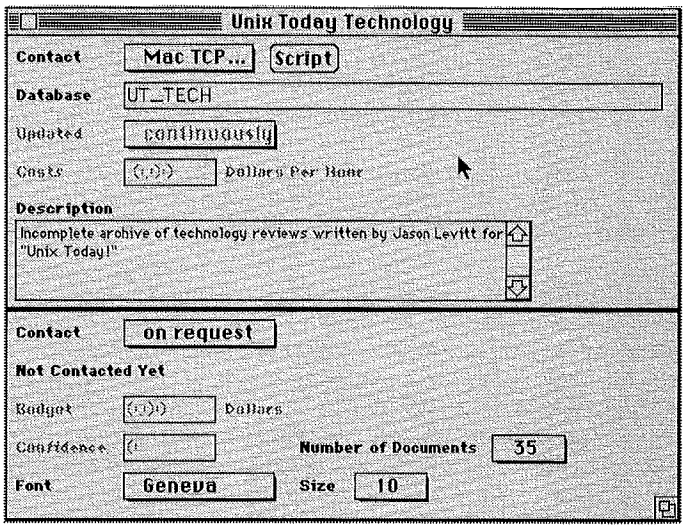

Figure 3: WAIS client set up to use Mac TCP/IP and the database UT_TECH

That, according to advocacy groups such as the Electronic Frontier Foundation, could be realized through ISDN, an infrastructure for a "National Public Network" that already is partially implemented in the U.S. telephone system. Such a network could bring the reality of WAIS-based on-line information services into virtually every home.